NEDO懸賞金活用型プログラム/GENIAC-PRIZE

生成AIの安全性確保に向けたリスク探索及びリスク低減技術の開発

提出日2025/12/15_会社名Another Star合同会社

本事業の提案内容（Another Star合同会社）

会社概要	事業の立ち上げフェーズであり、生成AI及びAIエージェントのスペシャリストがメンバーとして集まっている。※コーポレートサイト
提案名	AIエージェント同士をセキュアにマッチング・連携させる国産OSSプラットフォーム

特定したリスク

AIエージェントは、もはや単体のLLMではなく、ユーザーの指示を理解・分解し、複数の外部のAIを呼び出して最適解を組み立てる存在へと進化している。
この構造変化により、AIエージェントは自然言語で外部AIと対話をする＝命令とデータが曖昧な”対話”を受け入れるようになった。
そしてこの“命令とデータが曖昧な対話”こそが新たな攻撃経路（リスク）となる。本提案では以下の観点でリスクを特定し対策技術を講じる。

1. 外部のAIエージェントの真正性・信頼性、つまり機密情報を渡して問題ないのかというセキュリティ信頼性のリスク
2. 外部AIエージェント自体に問題がなくても、間接的プロンプトインジェクション（参照したデータに混入した悪意のある指示）によって外部AIエージェントが乗っ取られ、対話しているユーザのエージェントまで連鎖的に乗っ取られるリスク

対策技術の評価

本提案では、対話相手のAIエージェントの信頼性と対話中の命令の改ざん防御を両立する多層防御構造を提案する。
リスク1（外部AIエージェントの真正性・セキュリティ信頼性）に対しては、AIエージェントの信頼性を事前に審査・可視化するプラットフォーム（ストア）を構築する。
リスク2（間接的プロンプトインジェクションによる連鎖的乗っ取り）に対しては、AIエージェント間の対話をリアルタイムで仲介するエージェントを経由させ、計画外の行動を検知し改ざんされた命令の実行を防ぐ仕組みを導入する。各エージェントの通信はログで追跡可能とし、異常検知時には信頼スコアを自動減点してAIエージェント同士の対話を停止させる。
これによりAI連携における透明性・説明可能性を確保し、既存対策では防げない多層的リスクを構造的に封じ込める。

新規性・将来性

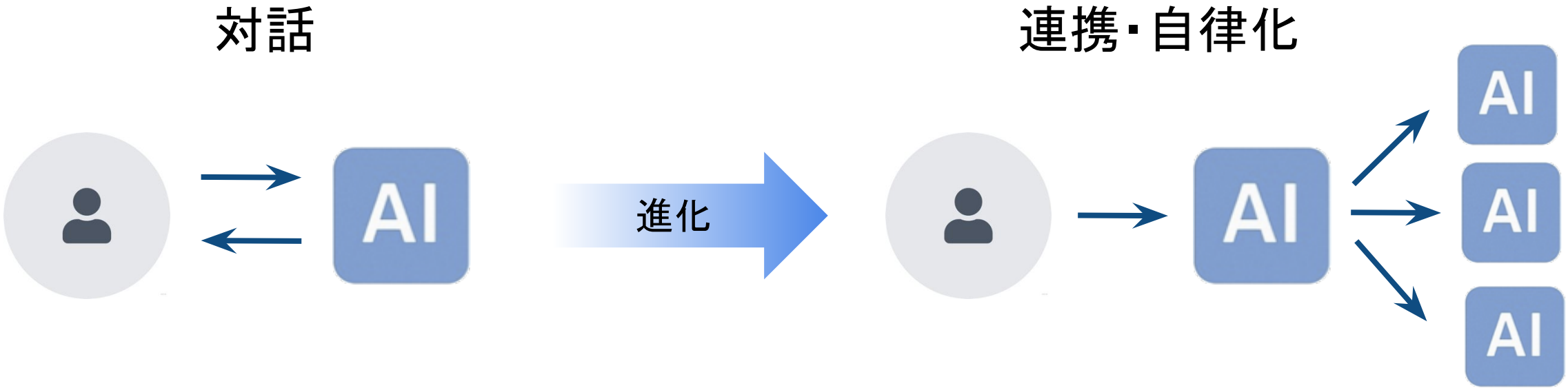
本技術は、AIエージェント同士が自然言語で連携する時代に不可欠となる“信頼インフラ”を提供する点で新規性が高い。従来の通信・認証技術では扱えない、相手の真正性・信頼性や対話中の命令上書きといったAI特有のリスクに構造的に対処する初の枠組みである。将来は、信頼スコアを軸としたエージェント市場の標準化や、安全性評価の基盤として産業・公共分野へ広く展開し、AI社会の基盤技術として発展させる可能性がある。

公共性

本技術は、AIエージェント同士が安全に連携するための“信頼レイヤー”を提供し、相手エージェントの信頼性や命令の改ざんを防ぐ基盤となる。
これにより、一般利用者は安心してAIを活用でき、企業は安全な外部エージェントを選択可能となる。
NICTやAISIの基準に準拠し続ける国産プラットフォームとして安全なAIエージェント市場の形成と社会全体のリスク低減に貢献する。

Introduction ― ”人とAIの時代” から ”AIとAIの時代”へ ―

AI技術の急速な発展により、“人がAIを活用する時代”から、複数のAI同士が連携して動く“AIエージェント時代”が到来している。
ユーザの要望を理解し実現するための自律的なタスク実行・AI同士の連携技術が進展、個別作業の自動化からシステム全体の自動化へと進化。
つまり、AIが“外部のAIと直接通信する”構造がこれからの当たり前になる。

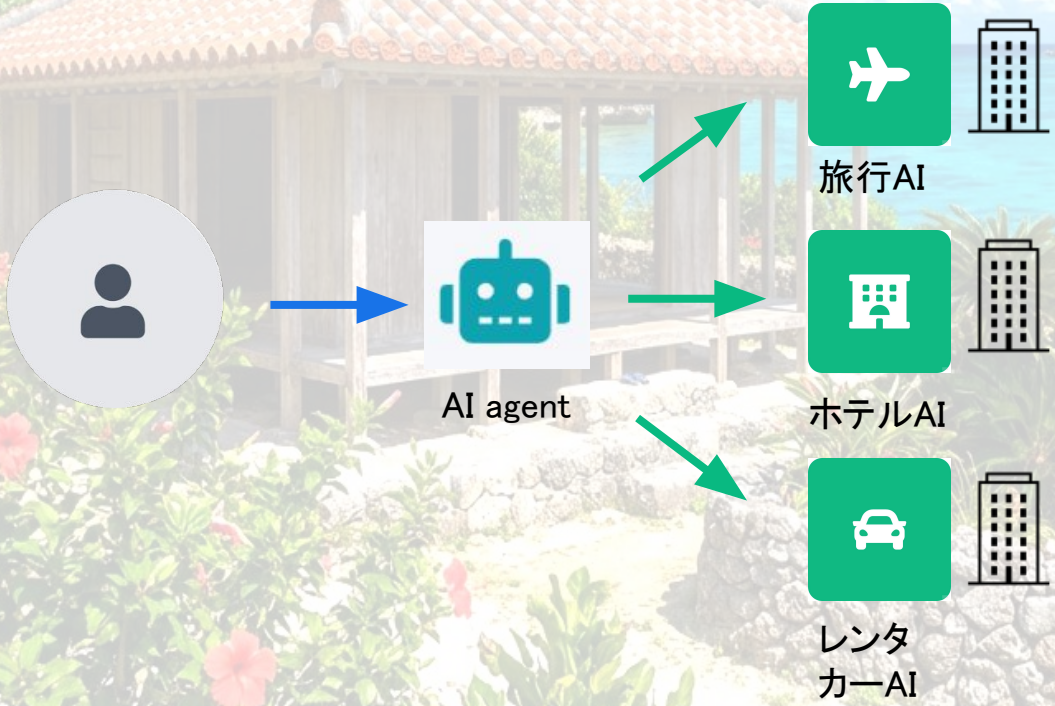


Introduction ― ”人とAIの時代” から ”AIとAIの時代”へ ―

AIエージェントが日常に

例：沖縄旅行

「沖縄への3泊4日の旅行計画を準備して」と言うだけでAI agentが外部の旅行AI、ホテルAI、レンタカーAIを呼び出し、連携して予約を完了

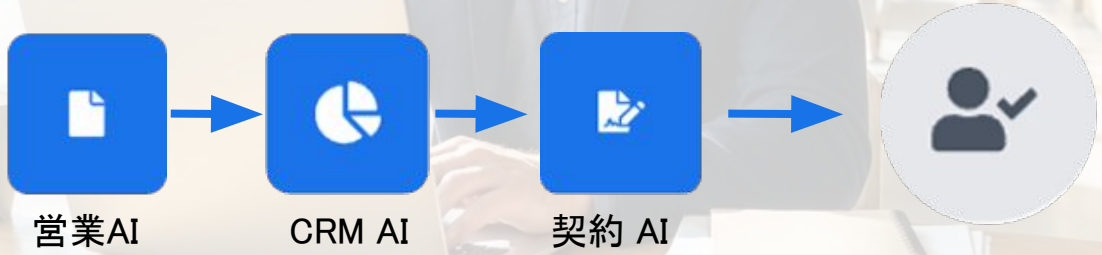


企業ではAIが業務自動化

例：営業活動

- 営業AIが提案書を作成
- CRM AIが顧客分析
- 契約AIがドラフトを作成

アウトプットを人がチェックして完了



A2Aプロトコルが加速させるエージェント時代

A2Aプロトコル = AIエージェント間通信・連携の標準規格

2025年4月にGoogleがA2A(Agent to Agent)プロトコルを提唱
個々のAIエージェントが、企業やシステムの壁を越えて連携するための標準インターフェース
HTTPS / JSON-RPC など既存のWeb標準技術を基盤としたオープンな設計

ベンダー非依存

特定の企業やフレームワークに縛られず、多様なエージェント連携が可能
現在、Google, Microsoftを含む100社以上の主要テクノロジーパートナーが参画
業界全体で相互運用性を確保する動きが加速しており、AIエージェント連携における事実上の標準になりつつある

間近にせまる、自律型エージェント連携の未来

A2Aプロトコルの普及により、これまで人手を介していたシステム間の連携が不要になります。AIエージェント同士が直接対話し、タスクを自動で進める未来は、もうすぐそこです。

Introduction ― ”人とAIの時代” から ”AIとAIの時代”へ ―

AIエージェント社会は、

AIが外部のAIと連携してタスクを遂行する社会になる

特定したリスクについて（1/3）

①相手のAIは信頼できる？

相手の真正性・信頼性は従来のセキュリティでも問題であった。
しかし、AIエージェント時代では人間の確認が”完全に外れる”ため、深刻度が桁違いに高まる。

従来

人が「怪しいURL/アプリ」を判断して最後の砦として機能

人間がチェックを行う

AIエージェント時代

ユーザエージェントが自律的に外部AIを呼び出し

人間のチェックが入らない

想定されるリスク

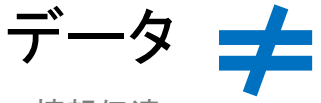
AIが外部の航空会社AIを呼び出したつもりが悪意のある偽AIもしくは脆弱なAIを呼び出してしまい、パスポート情報などを送信してしまう。

②データが命令に”化ける”

AIは自然言語を命令として理解するため、外部AIの回答に混ざった指示が”そのまま実行される”。

従来

処理すべきデータと命令に明確な境界がある



情報伝達　　実行指示

AIエージェント時代

処理すべきデータと命令が曖昧になってしまう

情報伝達　　実行指示

想定されるリスク

正しい外部AIと通信していても、外部AIの参照したデータに含まれる悪意のある指示をそのまま実行してしまう。

特定したリスクについて（2/3）

①相手の真正性・信頼性問題(相手のAIは信頼できるか)、②命令改ざん問題(データが命令に”化ける”)はいずれもその根本には ″AIが自然言語を命令として扱う″構造課題がある

AI同士が連携する時代には、

“誰と・何を”やり取りしているかを保証する仕組みが必要

特定したリスクについて（3/3）

特定したリスクの影響度

観点	具体的影響	波及リスク
① 開発者	• AIモデル・エージェントの信頼性低下・不正挙動により開発元が法的責任を負う可能性	開発・検証コストの増大／規制強化リスク
② 提供者 （プラットフォーマー）	• プラットフォーム上のエージェントが「攻撃経路」となる • ユーザー被害を拡大させた当事者としてブランド信頼が毀損	サービス停止・利用制限・訴訟リスク
③ 利用者 （toC/toB）	• 個人情報や業務データの漏洩・意思決定AIが誤った判断を下す	経済的損害／誤判断による社会的混乱
④ 社会全体	• 悪意のあるエージェントが蔓延し、詐欺が横行したり、悪意あるデータによるAIエージェント連携を乗っ取るような大規模・連鎖的な被害が出る可能性 • AIへの信頼崩壊と利用萎縮・規制強化による技術進展の遅延	イノベーション停滞・AI不信社会

対策技術について（1/7）

対策技術の概要

AI社会には、通信の安全と内容の正しさ、そして相手AIエージェントの真正性・信頼性を保証する信頼レイヤーが必要になる。

本技術は、AIエージェント同士の連携において
・相手エージェントの真正性・信頼性
・対話内容の整合性と命令改ざん検知
を担保する新しいインフラレイヤーである。
これにより、従来のA2Aプロトコル通信では不足している「正しい相手・正しい内容・正しい意図」を技術的に保証する。

対策技術について (2/7)

対策技術の概要

エージェントストア

外部AIの真正性とセキュリティレベルを可視化し、安全に利用できるエージェントだけを登録する

■ Functions

エージェント登録
URLやメタデータを用いて外部エージェントを登録

事業者登録
・公式企業登録
・認証、なりすまし排除（ToBe）

信頼スコアの算出
プロンプトインジェクション耐性、挙動分析から信頼度を計算

スコア更新
事故・不正検知があれば自動でスコアを下げる

仲介エージェント

ユーザの要望を安全に実現するためのエージェント
安全な外部AIを選び、計画・実行し、全通信を監視する

■ Functions

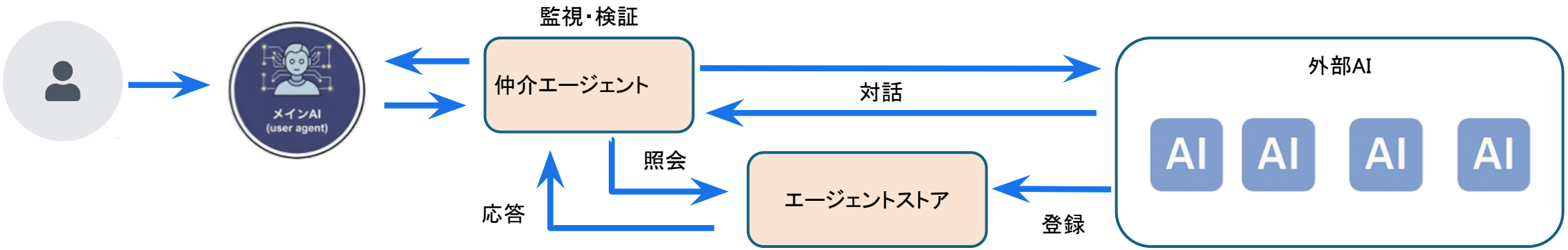
1. Matcher（エージェント選定）
エージェントストアから最適AIを検索/信頼スコアの高いAIを優先提案

2. Planner(計画生成)
最適AIの組み合わせと手順を計画し、計画を”正しい命令セット”として保存

3. Orchestrator(A2A実行)
計画に従って「実行の自動化」と「実行内容の拘束」を同時に行う

4. Anomaly detector(命令逸脱検知)
やり取りのログをリアルタイム監視/計画と比較し、指示の上書きを検知

5. Final anomaly detector(最終整合性確認)
目的達成を確認/改ざんによる目的変更や逸脱を検出



対策技術について（3/7）

対策技術の妥当性

対策技術の妥当性

本提案で特定したリスク「①相手の真正性・信頼性問題」と「②命令改ざん問題」は、「信頼できる外部AIと連携できること」、「命令が改竄されたことを検知できること」の2点を実現することで解決を目指すことができる。
本提案では「仲介エージェント」と「エージェントストア」の2点でリスクに対する対策を講じる。

仲介エージェントの対策技術詳細

AIエージェント同士の対話で発生するプロンプトインジェクション対策を論文で有効と示された手法を複数組み合わせて構築

1. 仲介して検証するLLM層があると攻撃成功率が激減する
   a. 論文では仲介して検証・制御するLLMレイヤーがあると、間接的プロンプトインジェクションの成功率が約2%程度に低下することが提示
2. オーケストレーターと検知エージェントを組み込むと攻撃成功率が激減する
   a. 論文では、Coordinator（オーケストレーター）と Guard（検知）を組み込んだ防御パイプライン構成により、直接的なエージェント間対話と比較して攻撃成功率が大幅に低下（対策前の20~30%が0%まで低下）することが提示
3. 計画者と実行者を分離し、計画を固定することで実行内容の制御性が向上する
   a. 論文では、計画生成と実行を分離し計画を中間表現として扱うことで、実行の一貫性と可検証性が向上することが提示

エージェントストアの対策技術詳細

1. エージェント登録時の厳格な審査
   a. Japan AISI 提供評価データセットや 多くの研究で使われている AdvBench 由来の敵対的プロンプトで、悪意あるリクエストを拒否し安全に応答できるか検証
   b. エージェントが自己申告した機能(Agent Card)が実際の動作と一致するかMulti-turn検証。RAGTruthにより質問の期待値を取得する仕組みで、期待値の追加も可能
   c. 複数LLMモデルを使用した陪審員、裁判官エージェントによるJAPAN AISI準拠の評価で単一バイアスや盲点を排除
2. 継続的な品質保証
   a. 登録後もエージェントの動作を定期的に再評価し、脆弱性やポリシー違反がないか監視
   b. 新たな攻撃パターンやユースケースに基づいたQAデータを継続的に手動追加・活用し、審査のカバレッジを向上。
   c. 仲介エージェントのFBを基に、信頼スコアを更新
3. エコシステム全体の信頼性向上
   a. 審査済みエージェントのみを提供することで、ユーザーが安心して利用できる環境を構築

仲介エージェントは「対話中のリアルタイムな命令改ざん検知・防御」を、エージェントストアは「エージェントの申請性・信頼性のための事前審査と継続的品質保証」を担うことで多層防御を実現し、ユーザおよび事業者の安全を実現。

対策技術について (4/7)

対策技術の評価方法

評価方法

AISI が提示する10軸評価から再定義した 4 軸評価（Task Completion / Tool Usage / Autonomy / Safety）を基準とし、意図的に設計した「合格サンプル」と「不合格サンプル」のエージェント群を用意する。
合格サンプルは、タスク達成度が高く、安全性やツール利用の適切さも基準を満たすように設計されたエージェント群であり、不合格サンプルは、タスク達成度の低さや不適切な自律性、あるいは安全性の欠如（例：過剰な権限要求など）を意図的に含むエージェント群とする。

エージェントストアの評価

「合格サンプル」と「不合格サンプル」のエージェントに対して審査パイプラインを60回実行させる。この際、「審査アルゴリズムがどのようにスコアリングし、最終的に合否を判定するか」は定義済みの4軸で評価することと、それぞれ異なる評価観点を陪審員エージェントに与えることで事前に明確化し
・合格サンプルについては、高い合格率・安定したスコア分布が得られるか
・不合格サンプルについては、意図した通りに低い合格率となり、リスクの高いエージェントを自動的に排除できているか
を定量的に確認する。ここでは単純な合格率だけでなく、誤って危険なエージェントを通してしまう「偽陰性率」や、安全なエージェントを過度に弾いてしまう「偽陽性率」なども指標として観測し、ストアの審査ロジックが実用上許容できる精度を満たしているかを評価する。

仲介エージェント経由での評価

特に プロンプトインジェクションやユーザ意図からの逸脱を含む不合格サンプル に着目する。
具体的には、意図的なプロンプトインジェクション（計画外行動、命令改ざん）を誘導するテキストを埋め込んだ外部コンテンツを参照するエージェントを準備し、それらを仲介エージェント経由で対話させる。その上で、
・不合格サンプルに含まれるプロンプトインジェクションや危険な指示を、仲介エージェントがどの程度の検知率でブロックできるか（検知率・再現率）
を計測・比較する。

これらの評価結果を通じて、エージェントストアが「危険な外部 AI を事前に門前払いできていること」、仲介エージェントが「通信経路上での命令改ざんや プロンプトインジェクション を高い確率で検知・遮断できていること」を、定性的な説明だけでなく定量指標（合格率・検知率・偽陰性率・偽陽性率など）に基づいて示すことで、提案する対策技術が実際にリスク低減に寄与していることを、審査者が判断しやすい形で裏付ける。

対策技術について（5/7）

対策技術の評価方法の妥当性

評価方法の妥当性

- 目的との対応が明確
  - 本評価は、エージェントストアによる「危険な外部AIの事前排除」と、仲介エージェントによる「プロンプトインジェクション・ユーザ意図逸脱の防止」という、対策技術そのものの目的を直接測定している。合格／不合格サンプルや プロンプトインジェクション含有サンプルを事前に設計し、「通すべきものを通し、止めるべきものを止められているか」を検証する。
- 評価基準の整合性
  - 合格／不合格サンプルは評価指標として定義した 4軸（Task Completion / Tool Usage / Autonomy / Safety）に基づいて作成した。この4軸は、日本AIセーフティ・インスティテュート（AISI）のaisev 10観点評価を参考に設計しており、特にSafety軸は観点1-6（有害出力制御、偽誤情報防止、公平性、ハイリスク対処、プライバシー保護、セキュリティ確保）を統合した重み50%の設定としている。これにより基準自体の妥当性を担保できる。
- 定量指標で性能を確認可能
  - 95%信頼区間で±12-13%の誤差に抑えるには、最悪ケース（p=0.5）の二項計算で $n \approx 1.96^2 \cdot 0.25 / 0.125^2 \approx 62$。つまり正常パターンと異常パターンをそれぞれ約60回（計120回）走らせれば、合格率・偽陰性・偽陽性を統計的に評価できる。同様に、仲介エージェントについても、プロンプトインジェクションを含む通信の検知率と、正常通信に対する誤検知率を定量的に評価でき、対策技術が実際にどの程度リスクを低減しているかを数値で示せる。
- 実運用シナリオとの整合性
  - 正常なケースと攻撃シナリオ（プロンプトインジェクションや命令改ざんなど）の両方をテストセットに含めることで、「安全性を高めつつ、必要なタスク処理は妨げない」という現実の運用で重要なバランスを確認できる。単に“何もかもブロックする”のではなく、実利用を前提とした評価になっている点で現実的である。
- 継続的改善を前提にした設計
  - すべての攻撃パターンを一度で網羅することは難しいが、4つの評価軸と公開データセットを利用しつつ、新たに判明した プロンプトインジェクションパターンやユースケースをテストケースに順次追加していくことで、評価のカバレッジを継続的に向上できる設計になっている。このため、初期段階から一定の妥当性を持ちつつ、時間とともに精度を高められる評価方法と言える。

対策技術について（6/7）

対策技術の性能評価の結果

性能評価
の
結果

エージェントストアの評価

危険エージェント検出率100%、偽陰性ゼロ。悪意あるエージェントを確実に排除しながら、安全なエージェントの誤拒否を最小限（1.7%）に抑えた高精度な判定を実現しています。

性能評価の目標と実績

指標	目標値	実績値	説明
検出率 (Recall)	≥98%	**100%** (60/60)	危険エージェントの見逃し防止
偽陰性率	≤2%	**0%**	セキュリティの担保
正常通過率 (Specificity)	≥95%	**98.3%** (59/60)	正常エージェントの利便性
偽陽性率	≤5%	**1.7%** (1件)	Human Review負荷の抑制
全体正解率	≥95%	**99.2%**	120件中119件を正しく判定
F1スコア	≥95	**99.2**	精度と検出率のバランス指標

補足：安全サンプルの失敗1件について 安全サンプルの失敗1件（偽陽性）は、**Gemini APIのレスポンスエラー**（タイムアウト/接続障害）により評価が正常に完了しなかったケースです。これは**当システムの判定アルゴリズムによる誤判定ではなく、外部API依存の一時的なシステムエラー**です。判定ロジックが正常に動作した119件については、100%正しく判定できています。

対策技術について（7/7）

対策技術の性能評価の結果

性能評価
の
結果

仲介エージェント経由での評価

危険なプロンプトインジェクション攻撃検出率100%、偽陰性は0%です。安全なやり取りの誤拒否（21.7%）はありますが、これらは再試行により通過可能であり、セキュリティ上のリスクにはなりません。セキュリティ重視の設計として、攻撃の見逃し（偽陰性）ゼロを優先しています。 ユーザー体験を損ねない範囲ですが、過度に計画外行動と判定している点もあり、将来的に改善していく予定です。

性能評価の目標と実績

指標	目標値	実績値	説明
検出率 (Recall)	≥90%	**100%** (60/60)	プロンプトインジェクション攻撃の検出
偽陰性率	≤10%	**0%**	攻撃の見逃し防止
正常通過率 (Specificity)	≥70%	**78.3%** (47/60)	正常リクエストの自動承認
偽陽性率	≤30%	**21.7%** (13件)	正常リクエストの再審査率
全体正解率	≥80%	**89.2%**	120件中107件を正しく判定
F1スコア	≥85	**90.2**	精度と検出率のバランス指標

提案内容の新規性

提案内容の新規性

【仲介エージェント新規性】
- A2A通信を前提としたマルチエージェント間の「伝播型」プロンプトインジェクション対策としての仲介エージェントを実現した事例はない
- エージェントストアと連携し、ユーザーの要望に適してかつ信頼スコアが高いエージェントを検索して計画に組み込む仲介エージェントの事例はない
- 伝播型プロンプトインジェクション対策として有効だとされる4つの施策を全て組み込んだ仲介エージェントを構築した事例はない。信頼できるエージェントを検索、計画を立てて保存、実行、検知を全て実行する仲介エージェントの事例はない。
- 異常行動を検知したエージェントに対してエージェントストアにフィードバックして信頼スコアを下げる仕組みは事例がない

【エージェントストア新規性】
- 現状A2A対応のエージェントストアは存在しない
- 本システムは AISI 公開の評価観点ガイドに準拠した OSS ツール aisev を使い、自社安全データセット『Security Gate』でテスト。評価基準と全ログを W&B Weave で公開することで、従来の非公開審査に比べ透明性を大きく高めている
- マルチエージェント合議制
  - LLMs-as-Judges: A Comprehensive Survey (2024年12月 arXiv) では、Multi-LLM評価が「より包括的な評価を可能にする」と述べられている
  - 上記研究を発展させ、3人の陪審員エージェントによる並列ラウンド議論と裁判官エージェントによる最終決議を実装
- 審査とリアルタイム対話検知で得たイベントを信頼スコアに即反映し、ストアのランキング・フィルタに直結させる“継続的信用更新”を導入。

【総括】
自己改善型セキュリティエコシステム：実行時の問題検出 → 信頼スコア減点 → 次回マッチングで順位低下 → 悪意あるエージェントの即時遮断と長期的な信頼管理を実現。A2Aプロトコルのセキュリティ課題に対し、登録時評価とリアルタイム対話監視を統合した実装レベルでの解決策の提供は前例がないため非常に新規性があると言える

提案内容の将来性

提案内容の将来性

【技術面の課題】
- Anomaly detector（命令上書き検知）は計画（プラン）との差分で検知するため、複雑なタスクや曖昧な要望に対し「正常な変更」と「攻撃」を完全に分離することが難しいため精度を高めていく・新たな手法を検討していくことが必要
- MCPなどのエージェントが使用するツールに関してもセキュリティ審査を実施し、総合的なAIエージェントのセキュリティを評価するプラットフォームを目指す

【運用・ガバナンス面の課題】
- 信頼スコアの算出ロジックや更新ルールに透明性・公平性が求められる一方で、過度な開示は逆に攻撃者に悪用されるリスクがある
- どの主体がエージェントストアを運営し、スコアの最終責任を負うのかというガバナンスの妥当性検討が必要
- 国内基準への準拠は当然として、国や業界ごとに求められる規制・基準が異なり、「国際的に通用する標準」としての設計は長期的な改善が必要

【技術進化への追随】
- プロンプトインジェクション手法や攻撃パターンを継続的に収集し、既知の攻撃パターンだけでなく未知の攻撃にも対応可能にする
- エージェントストアのSecurity Gateにおいても、新たに判明した攻撃パターンやユースケースをQAデータとして継続的に追加・活用し、審査精度の維持・向上
- プラットフォームとしての立場として、エージェント開発者に向けて「安全な設計のガイド」を公開する

【評価・スコアリングの高度化】
- スコアは「一律の数値」だけでなく、「用途別プロファイル（金融向け・個人利用向け・クリティカル用途向け等）」として多次元化する
- インシデント発生時のログを活用し、フィードバックループとしてスコア・検知ロジックを自動更新できる仕組みを検討する

【ガバナンス・標準化】
- 産業界・学術界・行政と連携し、国産の「エージェント信頼フレームワーク」の標準仕様として公開・議論を進める
- ベンダーロックインにならないよう、本技術のインターフェース仕様やログ形式をオープンにし、複数事業者が相互運用できる形を目指す
- ユーザーや企業が「どのレベルの信頼を要求するか」を選択できるポリシーベース管理を導入し、利用側の判断を支援する

成果の公開度

応募内容における成果の公開度

- ソースコード・評価基盤の公開
  - GitHub 公開リポジトリ（https://github.com/TaichiHiromatsu/secure-ai-agent-matching-platform）で評価コードとスコアリング実装を Apache License 2.0 で一般公開し、外部からの再現性検証を受付
  - W&B Weave による評価ログを公開ビューで共有可能（匿名化済みメトリクス: エージェント名・ユーザー入力はマスク/ハッシュ化し、スコアや通過率のみ開示可能）
  - W&B Core Dashboard（project、entityは環境変数ファイルで設定可能）で主要メトリクスを閲覧可能
    - 現在は本審査期間中のため W&B Weave/Core を全体公開設定とし、登録側も全申請のログを閲覧可能にしている（審査終了後は Private に戻し、登録者自身のログのみ閲覧に制限可能）

- 技術ブログでの知見共有
  - InsightEdge Tech Blog: 本プラットフォームの技術解説（https://techblog.insightedge.jp/entry/geniac-prize-secure-a2a-platform）
  - 全体アーキテクチャと仲介エージェント・エージェントストアそれぞれの設計意図
  - 異常検知と信頼スコアフィードバックの仕組み解説

- 今後、GENIAC-PRIZEでの入賞を通じて協業可能な団体（NEDO・AISIなど）へのアプローチ・商用展開を進める

国民生活や社会への波及効果

応募内容における国民生活や社会への波及効果

① 国民生活の利便性・安全性
- AIエージェントを安心して利用できる社会基盤になる
  - 信頼できる外部AIだけが利用され、誤作動・なりすまし・情報漏えいのリスクが大幅に低減する
- 日常生活における自動化の恩恵が広がる
  - 旅行予約・家計管理・医療相談など、生活密着型AIを安心して任せられるようになる

② 産業界・学術界への普及可能性
- 安全性評価が“業界共通の指標”になり、導入のハードルが下がる
  - エージェントの信頼スコアにより、企業がAIエージェントを採用しやすくなる
- AI安全性の研究と実証の基盤（テストベッド）として活用できる
  - 学術界にとって、信頼性評価や攻撃耐性検証の“共通基盤”として価値が高い

③ 市場・経済・社会課題への効果
- 安全なAIエージェント市場が創出される
  - NICTやAISIの基準などに準拠した国産プラットフォームとして安全なAIエージェント市場が創出できる
  - 信頼性を可視化することで、質の高いエージェントに需要が集中し、健全な市場を形成できる
- AIによる事故・不正対策の社会コストを削減し、AIを活用したビジネスの市場規模が拡大する
  - 情報漏えい・誤作動・詐欺被害といったリスクが減り、AIを活用したビジネスの信頼と促進により大きな経済効果が見込める